A Bayesian Morphometry Algorithm

Hanchuan Peng *, Edward Herskovits, and Christos Davatzikos

Center for Biomedical Image Computing,

Department of Radiology,

School of Medicine,

Johns Hopkins University,

601 N Caroline St, JHOC 3230,

Baltimore, MD, 21287, USA.

Tel: 410-955-7422

Fax: 410-614-3896

Email: phc@cbmv.jhu.edu,

ehh@braid.rad.jhu.edu,

hristos@rad.jhu.edu

This work was supported by The Human Brain Project, National Institutes of Health grant R01 AG13743, which is funded by the National Institute of Aging, the National Institute of Mental Health, the National Aeronautics and Space Administration, and the National Cancer Institute. Dr. Herskovits was also supported by a Richard S. Ross Clinician Scientist Award.

A Bayesian Morphometry Algorithm

Hanchuan Peng *, Edward Herskovits, and Christos Davatzikos

Center for Biomedical Image Computing,

Department of Radiology, School of Medicine, Johns Hopkins University,

601 N Caroline St, JHOC 3230, Baltimore, MD, 21287, USA.

Email: phc@cbmv.jhu.edu, ehh@braid.rad.jhu.edu, hristos@rad.jhu.edu

Abstract

Most methods for structure-function analysis in medical images usually are based on voxel-wise statistical tests performed on registered Magnetic Resonance (MR) images across subjects. A major drawback of such methods is the inability to accurately locate regions that manifest nonlinear associations with clinical variables. In this paper we propose Bayesian Morphological Analysis (BMA) methods, based on a Bayesian-network representation, for the analysis of MR brain images. First, we describe how Bayesian networks can represent probabilistic associations among voxels and clinical (functional) variables. Second, we present a model-selection framework, which generates a Bayesian network that captures structure-function relationships from MR brain images and functional variables. We demonstrate our methods in the context of determining associations between regional brain atrophy (as demonstrated on MR images of the brain), and functional deficits. We employ two data sets for this evaluation: the first contains MR images of 11 subjects, where associations between regional atrophy and a functional deficit are almost linear; the second data set contains MR images of the ventricles of 84 subjects, where the structure-function association is nonlinear. Our methods successfully identify voxel-wise morphological changes that are associated with functional deficits in both data sets, whereas standard statistical analysis (i. e., t-test and paired t-test) finds only some of these changes in the linear-association case, and fails in the nonlinear-association case.

Index Terms – Bayesian network, Bayesian morphological analysis, voxel-based morphometry, morphology-function analysis, atrophy detection, computational anatomy

1. Introduction

Voxel- and deformation-based morphometry have been increasingly used to identify morphological abnormalities, such as atrophy, without the need to define *a priori* specific regions of interest. Many different approaches [1-2, 9-17, 21, 32-35] have been proposed for generating statistical maps that identify groups of voxels that display differences in morphology, or voxels for which significant correlations exist among morphological and clinical measurements. Morphological measurements can be computed from the deformation field used to spatially normalize subjects into a stereotaxic space [13, 17, 32-33], from residual variability in the spatial distribution of gray and white matter after spatial normalization [1, 16], or from tissue-density maps obtained after mass-preserving spatial normalization [10,12].

For the purpose of morphology-function analysis, particularly voxel-wise morphometry, one of the first steps is warping MR images into a normalized space (i. e., registration), to ensure that voxel attributes across subjects can be compared. A widely used brain-image registration technique is the smooth parametric transformation [1, 14-15, 35], which is provided in the SPM99 software package (http://www.fil.ion.ucl.ac.uk/spm/spm99.html). Our group has previously developed another method referred to as Spatial Transformation Algorithm for Registration (STAR) [10, 37], which utilizes a high dimensional elastic transformation. Coupled with the high-dimensional elastic transformation is a procedure that preserves information about the volumes of different anatomical structures, by constructing tissue-density maps, in which relatively higher density at a particular structure implies that this structure has a relatively higher volume prior to spatial normalization. This procedure is a key component of our approach, since spatial normalization changes the anatomy of individual subjects, by making each subject's anatomy similar to that of a template. Therefore, having a mechanism that preserves volumetric information during

spatial normalization is critical.

Regardless of the type of morphological variables being considered, e.g. voxels or regions, most existing morphology-function analysis methods rely on voxel-wise linear statistics, such as t-tests (TT), paired t-tests (PT), and ANOVAs. Such statistics compare only the means and variances of variables among different groups; therefore, methods based on these statistics may not be able to detect nonlinear morphology-function associations. Second, even for linear associations, these methods usually require a predefined confidence interval, or p-value threshold, to generate regions of interest (ROIs). Third, these statistical tests generally do not directly describe the relationships among the generated ROIs. Other methods, such as principal-component analysis and partial-least-square analysis [26], can be expected to capture some image-behavior relationships. However, few of these methods can bring together many conditional probabilities of these variables to give a statistical conclusion. In particular, we distinguish between linear associations among continuous variables, which can be evaluated using methods based on the general linear model (GLM), such as ANOVA or linear regression, and nonlinear associations among continuous or categorical variables, which may not be captured by a GLM.

In this paper we use Bayesian networks (BNs) [19, 22, 25] to represent the probabilistic associations between MR image voxels and clinical variables. A BN is a Directed Acyclic Graph (DAG) model describing the probabilistic relationships among variables; each node represents a variable, and directed edges coming into a child node indicate that there is a corresponding conditional-probability distribution for the child, given the joint states of its parents. Each node without parents is associated with a prior probability distribution. In this framework, voxel-morphology variables (e. g., dilated, contracted) and clinical variables are nodes, and morphology-function analysis is equivalent to the generation of a Bayesian network from MR image data and clinical information for each subject. The Bayesian network techniques needed in this paper are formalized below in Section 1.A.

We evaluate our methods by trying to detect cerebral atrophy in structural MR brain images, in the setting of changes in clinical variables. Cerebral atrophy [20] is a degenerative process that generally occurs after 55 years of age, although it may occur much more rapidly in certain diseases. In this process, the brain loses mass and volume, causing the cerebral ventricles to dilate. Many cortical, subcortical, and mixed cortical-subcortical encephalopathies, such as Alzheimer’s disease and Parkinson’s disease, have atrophy as their primary structural

manifestation. This application is a typical example of morphology-function analysis, where significant associations between brain morphological changes and clinical variables, such as aphasia or apraxia, are of interest.

This paper is arranged as follows. Section 1.A briefly introduces Bayesian networks. Section 2 describes an overview of our approach, describes a framework for finding the structure of a Bayesian network from image (i. e., voxel) and clinical data, and further presents two methods implemented within this framework for generating sets of equivalent voxels. Section 3 describes our performance metrics. Sections 4 and 5 illustrate experimental results on a linear-association data set of cerebral MR images, and on a nonlinear-association data set of ventricular MR images, respectively. After discussion in Section 6, we present our conclusions in Section 7.

A. Bayesian Networks

Suppose we have n nodes $X=\{x_1,\ldots,x_n\}$; a Bayesian network for X consists of a DAG structure $\mathbf{S}$ and a set of local distribution functions $p(x_i|\pi_i,\theta_\mathbf{S},\mathbf{S})$, where π_i is x_i's parent node set and $\theta_\mathbf{S}$ is the parameter set of all conditional probabilities. The structure $\mathbf{S}$ encodes the independence statement that $p(X \mid \theta_\mathbf{S},\mathbf{S}) = \prod_{i=1}^{n} p(x_i \mid \pi_i,\theta_\mathbf{S},\mathbf{S})$; that is, the structure of a Bayesian network defines a decomposition of a joint distribution into the product of conditional-probability distributions, based on the notion of conditional independence, which we elaborate below. Many model-selection algorithms (see [22-23, 38] for good reviews) have been proposed to construct Bayesian networks from data. Often these algorithms are based on assumptions similar to the following [8, 22-24, 31, 3]:

1. Each variable is discrete, having a finite number of states. We use x_i^k and π_i^j to denote the kth state of x_i and the jth possible joint configuration of π_i, respectively. We use r_i and q_i to denote the number of possible states of x_i and the number of possible joint configurations of π_i, respectively.
2. Each local distribution function $p(x_i|\pi_i,\theta_\mathbf{S},\mathbf{S})$ consists of a set of distributions defined as the parameters

$$\theta_{ijk} \equiv p(x_i^k \mid \pi_i^j,\boldsymbol{\theta}_\mathbf{S},\mathbf{S}), \tag{1}$$

where for all i, j, k, $\theta_{ijk} > 0$ and $\sum_{k=1}^{r_i} \theta_{ijk} = 1$. Denote the parameter set $\theta_{ij} = \{\theta_{ij1}, \cdots, \theta_{ijr_i}\}$.

3. The parameter sets θ_{ij} are mutually independent, so that $p(\boldsymbol{\theta}_S \mid \mathbf{S}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} p(\boldsymbol{\theta}_{ij} \mid \mathbf{S})$.

4. Each parameter set θ_{ij} has a Dirichlet distribution, giving $p(\boldsymbol{\theta}_{ij} \mid \mathbf{S}) = \mathrm{Dir}(\boldsymbol{\theta}_{ij} \mid \alpha_{ij1}, \cdots, \alpha_{ijr_i}) \propto \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}-1}$,

   where hyperparameters $\alpha_{ijk} > 0$ for every i, j, k.

5. The data set **D** is complete, that is, every variable is observed in every case of **D**.

Under these assumptions, the parameters remain independent given **D**:

$$p(\boldsymbol{\theta}_S \mid \mathbf{D}, \mathbf{S}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} p(\boldsymbol{\theta}_{ij} \mid \mathbf{D}, \mathbf{S}), \tag{2}$$

and the posterior distribution of each θ_{ij} has the Dirichlet distribution

$$p(\boldsymbol{\theta}_{ij} \mid \mathbf{D}, \mathbf{S}) = \mathrm{Dir}(\boldsymbol{\theta}_{ij} \mid \alpha_{ij1} + N_{ij1}, , \cdots, \alpha_{ijr_i} + N_{ijr_i}) \propto \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk} + N_{ijk} - 1}, \tag{3}$$

where N_{ijk} is the number of cases in **D** in which $x_i = x_i^k$ and $\pi_i = \pi_i^j$. $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ is the total number of cases in which π_i assumes the jth joint parent configuration.

The marginal likelihood was first derived in [8, 24] as

$$p(\mathbf{D} \mid \mathbf{S}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(\alpha_{ij} - 1)!}{(N_{ij} + \alpha_{ij} - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \tag{4}$$

where $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$. Since α_{ijk} is often chosen as 1, $\alpha_{ij} = r_i$ and

$$p(\mathbf{D} \mid \mathbf{S}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \tag{5}$$

Among many Bayesian network learning methods, we have particular interest in K2 [8, 24] . K2 uses the metric eqn.(5) to compute the conditional probability of a candidate Bayesian-network structure **S** (i. e., the associations among the variables), given the data **D**. With uniform priors over possible network structures, the

Bayesian metric is proportional to the likelihood in eqn.(5). Because the number of possible network structures is exponential in the number of variables, it is impossible to completely evaluate all possible network structures to find the best one [24]. In fact, for the purpose of voxel-wise morphometry, there are hundreds of thousands of voxel variables, so it is impractical to directly apply the K2 algorithm to morphology-function analysis. As we describe later, we solve this problem by searching a specific subset of possible network structures; this subset is suitable for capturing associations among voxels and clinical variables, and for finding sets of equivalent voxel variables.

A basic concept regarding Bayesian networks is d-separation, which is defined as follows [25]: two variables *v* and *u* in a Bayesian network are d-separated if, for all paths between *v* and *u* there is an intermediate variable *w* such that either the connection is serial or diverging and the state of *w* is known, or the connection is converging and neither *w* nor any of *w*'s child nodes have known states. This blocking of evidence transmission is reflected as conditional independence among variables. Variable *u* is conditionally independent of *v* given variable *w* if $p(u \mid v, w) = p(u \mid w)$. In this case, knowledge of *v* will not alter the probability of *u*. If *w* is empty, we say that *u* and *v* are marginally independent. The conditional independence appears in Bayesian network paths of serial and diverging connections. In this paper we utilize d-separation to find candidate sets of equivalent variables.

Latent-variable induction in Bayesian network models has been presented as a clustering method [4, 6-7]. In this paper we use latent-variable induction to generate sets of equivalent variables from the above-mentioned candidate sets.

2. Bayesian Morphological Analysis

In this section we propose a Bayesian Morphological Analysis (BMA) algorithm, which detects morphology-function associations between brain morphological variations and clinical variables. Like other morphometry methods, BMA requires a preprocessing stage for the images. BMA is based on the Bayesian metric of eqn.(5) and a heuristic model-selection method to generate a Bayesian network structure from the data. Next,

BMA produces sets of equivalent voxels based on Bayesian thresholding or Bayesian clustering.

A. Image Data Preprocessing

The purpose of data preprocessing is to generate the image portion of data **D** for the BMA algorithm. For purposes of illustration, most of our development will be presented in the context of finding associations between longitudinal brain atrophy and functional variable that might be associated with such atrophy. Suppose we have longitudinal MR images (at times t_1 and t_2) of a group of subjects (subject 1, 2, …), along with measurements of a categorical clinical variable, which could reflect performance in a neuropsychological battery of tests. The three major image-processing steps in the BMA framework, i.e. registration, subtraction and thresholding, are shown in Fig.1:

Fig.1 Three major image preprocessing steps, i.e. registration, subtraction and thresholding (binarization)

(1) *Registration:* In the registration step, brain images of different sizes and shapes are warped to a stereotaxic canonical space. In this paper we choose STAR [10, 21], which is a high-dimensional elastic-registration method. Unavoidably, registration introduces a complication, namely that it changes the morphology of an individual's brain. Therefore, it would be pointless to examine the morphology of spatially normalized brain images in a structure-function correlative analysis. In order to overcome this problem, we use an approach referred to as Regional Analysis of Volumes Embedded in Stereotaxic Space (RAVENS), which is described in detail in [12,10,21]. In this approach, 3D density maps for each tissue class, such as gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF), are generated separately. For example, assume that an individual brain has, due to atrophy, larger ventricles than those in the canonical-space template. Then the density of the corresponding CSF map will be high after spatial normalization, reflecting the fact that a relatively larger volume of CSF is forced to fit in a relatively smaller space. More generally, RAVENS maps reflect the regional volumetric structure of the brain of

each subject, with the tissue density of any structure being proportional to the actual volume of the structure in an individual brain, prior to spatial normalization. Since these maps are registered and reside in the same canonical space, they can be overlaid and analyzed on a voxel-wise basis.

(2) *Subtraction:* We subtract the pair-wise RAVENS maps to generate a difference image for each subject. Due to volume contraction, the atrophic regions of t_2 images will, on average, have lower intensity than the corresponding regions in t_1 images, since image intensity of the RAVENS maps reflects tissue density. As a result, the difference maps ($t_1 - t_2$) will generally have positive values in these regions, and negative values for regions that dilate over time, such as ventricles in the setting of progressive cerebral atrophy.

(3) *Thresholding:* Our current method applies only to categorical variables. Therefore, we binarize the longitudinal atrophy maps by thresholding them at zero. That is, any voxel value larger than 0 is set as state 1 (for "volume contraction") otherwise is set as state 0 (for "no volume contraction"). In regions of atrophy, binary-map voxels will in general assume state 1. At all other locations the binary maps would approximately have equal possibility to assume state 1 or state 0. The binary maps are used as the image data input for BMA.

In a binary map, we denote the entire brain as R_b and the (possibly discontiguous) region of atrophy as R_a. It follows that the non-atrophic brain region is

$$R_{\bar{a}} = R_b \setminus R_a .$$

B. Generating a Bayesian Network Structure From Data

In the BMA framework, we employ the following terms and notations:

- Pair-wise longitudinal MR images: a pair of MR images of the same subject obtained at different times, t_1 and t_2.
- Binary variable: a two-state variable; we label these states as 0 and 1.
- Voxel Variable (VV) *v*: a binary variable defined for each image voxel. A set of voxel variables is called a

VV-set and denoted as *V*.

- Functional Variable (FV) *f*: a discrete (not necessarily binary) clinical variable. For example, it could indicate whether a subject has a functional deficit (1) or not (0).
- Data **D**: a set of *N* binary maps (each has size $d_1 \times d_2 \times d_3$ for the three spatial dimensions) and *N* values for the FV.
- Association: a probabilistic association among a VV and an FV.
- Associated Variable (AV): a VV that is associated with a FV. If a VV is not an AV, we label it as a non-associated Variable (NV). The set of non-associated variables is the NV-set.
- Representative association Variable (RV) *u*: a VV that represents a group of AVs of similar associations, i.e., they have similar conditional distributions with respect to one or more FVs. The representative association variable set is called RV-set and denoted as *U*.
- Equivalent-variable set (e-set): the variable set represented by an RV. An e-set can be imagined as a (possibly discontiguous) region within which morphological changes affect volume loss in a homogeneous way.
- Candidate equivalent-variable set (c-set): the VV-set containing candidate equivalent VVs with respect to an RV.
- Bayesian network structure **S**: a Bayesian network consists of RVs and the functional variable(s). **S** is constructed based on the given data **D**.

In BMA, a Bayesian network **S** is constructed from **D** to achieve two goals: (1) identify all AVs from the complete VV-set *V*, and (2) classify all AVs into several subsets. AVs within a subset have similar conditional probability distributions with respect to a given FV, and AVs in different subsets have different conditional probability distributions with respect to a given FV. We accomplish both goals by finding the AVs, identifying RVs, and obtaining the corresponding e-set for each RV. Three major issues, i.e. the subset of possible Bayesian-network structures considered, the metric used to compare candidate network structures, and the search strategy, are next considered.

Fig.2 The general Bayesian-network structure for representing morphology-function associations

We propose using the general network structure in Fig.2 to represent associations among VVs and a FV. With this structure we suppose each u_i possesses an association with f, while each pair u_i, u_j (i,j=1,2,3… and $i \neq j$) are independent (i. e., there is no edge between u_i and u_j). Each u_i may represent a different kind of association between a VV and a FV. Hence this network structure is able to capture complex associations among these variables. Typically, for each u_i, its e-set contains more than one VV, since we expect voxels in the same group to have similar associations with a given FV.

We use the Bayesian metric [8, 24] $M(\mathbf{S})$, which is the conditional probability of network structure **S** given the data **D**, to evaluate candidate Bayesian-network structures. $M(\mathbf{S})$ has the following form:

$$M(\mathbf{S}) = p(\mathbf{S} \mid \mathbf{D}) = \frac{p(\mathbf{D} \mid \mathbf{S})\, p(\mathbf{S})}{p(\mathbf{D})},$$

where p(**D**|**S**) is computed using eqn.(5).

A larger Bayesian metric value indicates a larger probability that the corresponding Bayesian-network structure could have generated the observed data. Since we have no *a priori* preference regarding network structures, we assume the prior $p(\mathbf{S})$ is uniform, as in [8, 24]. Furthermore, because the prior probability of observing data **D** is a constant, $M(\mathbf{S})$ is proportional to the likelihood function, i.e. $M(\mathbf{S}) \propto p(\mathbf{D} \mid \mathbf{S})$, which takes the form of eqn.(5) for discrete variables of complete data **D**.

For computational purposes, we redefine the metric M as the logarithm of eqn.(5):

$$\begin{aligned} M(\mathbf{S}) &= \ln p(\mathbf{D} \mid \mathbf{S}) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{q_i} \{\ln[(r_i - 1)!] - \ln[N_{ij} + r_i - 1)!]\} \sum_{k=1}^{r_i} \ln(N_{ijk}!) \\ &\equiv -\sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{h=r_i}^{N_{ij}+r_i-1} \ln h \sum_{k=1}^{r_i} \sum_{t=2}^{N_{ijk}} \ln t \end{aligned} \tag{6}$$

C. Model Selection

Here we assume that it is unnecessary to find all associations among VVs and an FV in the network structure (see discussion at the end of this subsection). Instead, we wish to determine whether there is an association between an individual v_i and f, and whether v_i is conditionally independent of that f given the existing RVs. We propose the method shown in Fig.3 to generate the set of all RVs, and their corresponding e-sets.

Fig.3 Flowchart of the algorithm for RV-set generation.

The procedure in Fig.3 can be understood step by step as follows. Our algorithm first compares pairs of Bayesian-network structures with and without an edge from the current v_i to f (shown as Step 1 in Fig.3 and plots (a)(b) in Fig.4, where the five structures are denoted as $\mathbf{S}_a$, $\mathbf{S}_b$, $\mathbf{S}_c$, $\mathbf{S}_d$, $\mathbf{S}_e$, respectively). In particular, to find the variable that has the strongest association with f, the algorithm compares each pair of Bayesian networks shown in Fig.4 (a) and (b) over all v_i by computing the difference metric

$$dM(\mathbf{S}_a, \mathbf{S}_b) = M(\mathbf{S}_a) - M(\mathbf{S}_b) \tag{7}$$

Since dM is a function of v_i, we denote it as dM_i for convenience (as used in Step 1 in Fig.3). Thus, we obtain the maximum dM^* and the corresponding voxel v^*,

$$dM^* = \max dM_i \tag{8}$$

$$v^* = \arg\max_{v_i} dM_i . \tag{9}$$

If there are several variables whose difference metric values are equal, we choose the variable with the maximum $M(v_i)$ as v^*. If there are still several variables with the same metric value, we arbitrarily choose one of them as v^*.

If dM^* is not larger than 0, the data do not favor an edge from any VV to f, hence our algorithm stops (this condition is equivalent to the judgment subsequent to Step 1 in Fig.3). Otherwise we denote the variable v^* as u_1.

Then we exclude u_1 from the current VV-set and put all variables for which $dM_i \le 0$ in the current NV-set.

However we do not exclude these NV-set variables from VV-set because they might have higher order associations with *f*, although in the current order they are not linearly associated with *f*.

Each of the variables that render $dM_i > 0$ will have an edge to *f*. Hence we calculate the difference metrics of Fig.4(c) and (d) (i.e. compare models $\mathbf{S}_c$ and $\mathbf{S}_d$) for all VVs in the current VV-set,

$$dM(\mathbf{S}_c, \mathbf{S}_d) = M(\mathbf{S}_c) - M(\mathbf{S}_d) \tag{10}$$

If a variable v_i makes eqn.(10) less than or equal to 0, the data favor no edge conditioned on the existence of u_1. In that case, from the definition of conditional independence, we know that v_i and *f* are conditionally independent given u_1. The relationship between conditional independence and d-separation implies that the data favor the model $\mathbf{S}_e$ in Fig.4 (e) rather than $\mathbf{S}_c$ in Fig.4 (c). Therefore, it is reasonable to place the current NV-set as the c-set of u_1, and to assert that there probably are strong associations between these NV-set variables and u_1. The e-set of u_1 will be found in the c-set using the methods in next subsections. Intuitively speaking, the method to obtain the c-set is analogous to partial correlation [39], which is used to determine whether there is any linear relationship between two variables given another (or another group of) controlling variable(s).

The variables for which eqn.(10) is larger than 0 add more information even given the existence of u_1. Hence we reuse eqns.(8) and (9) to find the v^* that has the maximum difference metric of eqn.(10). Denote this v^* as u_2. Note that u_2 is typically independent of u_1, although this is not necessarily true.

We then exclude u_2 and the e-set of u_1 from the current VV-set (Step 3 in Fig.3). If the VV-set is non-empty, the algorithm continues, to generate the e-set of u_2, and the third representative variable u_3. This procedure is repeated until no variable remains in VV-set (the judgment subsequent to Step 3 in Fig.3). As shown in Fig.3, these iterations will generate a set of representative variables $U=\{u_i\}$ and their corresponding e-sets.

Fig.4 Alternative structure-function Bayesian-network models

We assume above that we need not compare all possible Bayesian networks to generate an adequate network structure. Regardless of whether this assumption is correct, the model-selection algorithm will produce the correct

u_1 for model $\mathbf{S}_a$. However, this assumption is probably incorrect for networks with more than one RV. For example, for structure $\mathbf{S}_c$ it is not impossible that the combination of u_2 and another AV could yield a larger metric value than the combination of RVs u_1 and u_2. One method to decrease the likelihood of such a problem is: first, obtain u_2 with the heuristic method in Fig.3, then replace u_1 in models $\mathbf{S}_c$ and $\mathbf{S}_d$ with other currently available AVs, to ensure that the combination of u_1 and u_2 will produce the largest metric value. We apply the same approach for structures with more nodes.

Another challenge is that the structure might not be sufficiently complex to represent associations among RVs and an FV. For example, it is not impossible that an additional edge from u_2 to u_1 in structure $\mathbf{S}_c$ could lead to a larger metric value. It appears that the only way to overcome this problem is to carry out an exhaustive search for all possible associations, which has exponential computational complexity in the number of variables. Nonetheless, because we need only to identify associations between AVs and the FV, the BMA algorithm can succeed even if we do not compare associations among RVs.

D. Bayesian Thresholding

In this paper we employ the concept of "probabilistic equivalence": two variables v and u are probabilistically equivalent if v and u have the same number of states, and $p(v \mid u) \cong p(u \mid v) \cong 1$ for each state.

The equivalent variables u_i are only searched for in the corresponding c-set (in our implementation, the c-set is set to the union of the c-set and u_i). An alternative method is to search all variables in the current VV-set, rather than only the c-set. However, we prefer searching the c-set, because a variable that would add more information would generally have a different association with u_i and could therefore be excluded from consideration.

A straightforward method to find the e-set of u_i is Bayesian thresholding (BT), which determines whether the association between each c-set variable and u_i is strong. In BT, two "equivalent" binary variables u and v are required to satisfy the following conditions:

$$
\begin{aligned}
&p(v=1\,|\,u=1)\approx p(u=1\,|\,v=1)\approx 1\geq p_{\mathrm{BT}}\\
&p(v=0\,|\,u=1)\approx p(u=0\,|\,v=1)\approx 0<1-p_{\mathrm{BT}}\\
&p(v=0\,|\,u=0)\approx p(u=0\,|\,v=0)\approx 1\geq p_{\mathrm{BT}}\\
&p(v=1\,|\,u=0)\approx p(u=1\,|\,v=0)\approx 0<1-p_{\mathrm{BT}}\ ,
\end{aligned}
\tag{11}
$$

where p_{BT} (<1) is a predefined threshold. An implementation of BT is shown in Fig.5, where the edges are indicated by the horizontal edge in Fig.4(e). Clearly, the threshold p_{BT} should be larger than 0.5, otherwise there could be no edges from any of the c-set variables to u_i.

The structure in Fig.5 does not necessarily mean that there is no association among v_j, v_{j+1}, …. On the contrary, typically there would be a fully connected Bayesian network for these equivalent variables. Therefore we emphasize that the network in Fig.5 is very different from the structure in Fig.2, where each u_i is supposed to be independent, or at least associated with a given FV in a different manner than the other u_j, and therefore will add more information to the existing structure.

Fig.5 The general Bayesian-network structure for the BT method

E. Bayesian Clustering

The major drawback of the BT method (and other threshold-based methods) is its reliance on a predefined threshold. We therefore developed another BMA algorithm, based on latent-variable induction in Bayesian networks, to cluster the c-set and obtain the e-set. In this approach, we transpose **D**, i. e., we consider a pseudo-variable set **C**, where each c_i is the variable representing all *i*th cases of every variable in the c-set, and at the same time, we regard the original variables as the pseudo-cases. The pseudo-data are denoted as $\mathbf{D}^{\mathbf{T}}$. Fig.6 shows this scheme, where L is an r_L-state latent variable with edges to each of the pseudo-variables c_i, i=1,2…,N. Each state of L corresponds to a set of pseudo-cases. The joint distribution of c_i is given in the following multinomial form:

$$p(C)=\sum_{j=1}^{r_L} p(L=L^j)\prod_{i=1}^{N} p(c_i)\,p(L=L^j)$$

The clustering method is to assume a number of states for the latent variable L, then estimate the unobservable cases of L. An approximation method [6], based on the Laplace-approximation and Bayesian information criterion, minimum description length, or Cheeseman-Stutz approximation, can be used. An alternative method is the Monte-Carlo approximation, which will produce a more accurate result if given enough time [6]. In this paper we consider one specific Monte-Carlo method, the Gibbs sampler [18], because we want the clustering results to be precise, while reducing the computational burden.

Fig.6 The latent-variable Bayesian-network structure for the BC method

In the Gibbs-sampler approximation, we first randomly initialize the unobservable pseudo-cases of L (with the assumed number of states r_L). Then we sequentially un-assign every latent pseudo-case and calculate its probability for each possible state, given the other pseudo-cases:

$$p(L_j = L^i \mid \mathbf{D^T} \setminus L_j, \mathbf{S}) = \frac{p(L_j = L^i, \mathbf{D^T} \setminus L_j \mid \mathbf{S})}{\sum_{k=1}^{r_L} p(L_j = L^k, \mathbf{D^T} \setminus L_j \mid \mathbf{S})}, \tag{12}$$

where $\mathbf{D^T} \setminus L_j$ denotes the data set $\mathbf{D^T}$ with the jth pseudo-case of L, i.e. L_j, removed; and the sum in the denominator runs over all possible states of L_j. Since both the numerator and denominator are probabilities that are computed based on an assumption of complete data, they can be computed using eqn.(5). Then, the results of eqn.(12) are used to sample a new pseudo-case. Next, all unobserved data can be reassigned to produce the new data $\mathbf{D^T}$. We iterate this procedure until the distribution of model parameters in Fig.6, which takes the form of eqn.(2), converges. We can rewrite such an indexing function as:

$$p(\boldsymbol{\theta}_S \mid \mathbf{D^T}, \mathbf{S}) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} p(\boldsymbol{\theta}_{ij} \mid \mathbf{D^T}, \mathbf{S}) \tag{13}$$

To calculate eqn.(13), we use eqns.(3) and (1). After the indexing function converges, typically most of the unobserved cases will not change states. Then we can find the pseudo-cases for which the real variables assume the same state with the RV u_i. These same-label variables are regarded as equivalent variables.

We call this clustering method Bayesian Clustering (BC); it has also been referred to as the candidate method [5-6, 30]. In [6], the authors were interested primarily in the indexing function value in eqn.(13) as opposed to the clustering results; this value is used only to decide when the algorithm should stop. In the setting of morphology-function analysis, we are primarily concerned about whether the clustering results are meaningful. Further, unlike [6], we neither select any particular fixed parameters nor use the normalized parameters.

The computational bottleneck of BC is eqn.(12). Fortunately, we can simplify it as the relative value of eqn.(5), which removes a lot of unnecessary computation of Bayesian metrics over the whole Bayesian network in Fig.6. In our implementation, this optimization makes the Monte-Carlo method very fast and is no longer a computational obstacle, especially when there are a large number of variables (pseudo-cases) in a c-set.

The parameter r_L, i.e. the number of clusters, must be set; however, it is not difficult to choose this parameter. First, when r_L is larger than a critical value r_{Lr}, which is the real number of clusters, the indexing function, eqn.(13), will typically converge to a value that is independent of r_L. Second, when $r_L > r_{Lr}$, typically there are only r_{Lr} different labels in the converged pseudo-cases of the latent variable. Therefore, we can simply set r_L to a large value, say, 6 or 7, even when we only expect to find 2 or 3 clusters. The only drawback of beginning with a large r_L is greater computational complexity, which is however not critical in practice.

3. Performance Metrics

Let us denote the detection results as R_r; this set is the union of the RV-set and all corresponding e-sets. To measure the performance of the algorithm, it is natural to define the Signal-Detection Rate (SDR) and Signal-Noise Ratio (SNR) as follows,

$$SDR = \frac{\Omega(R_r \cap R_a)}{\Omega(R_a)} \tag{14}$$

$$SNR = \frac{\Omega(R_r \cap R_a)}{\Omega(R_r \cap (R_b \setminus R_a))}, \tag{15}$$

where $\Omega(.)$ is the operator to calculate a region's volume, i. e., the number of voxels in that region. SDR measures

the fraction of detected ground-truth signal (and is therefore in the range [0, 1]), and SNR indicates the degree of false-positive detection. When the ground truth R_a is completely covered by R_b, eqn.(15) can also be written as

$$SNR = \frac{\Omega(R_r \cap R_a)}{\Omega(R_r \setminus (R_r \cap R_a))} = \frac{\Omega(R_r \cap R_a)}{\Omega(R_r) - \Omega(R_r \cap R_a)}$$

We expect that SNR may be larger than 1. Nonetheless, when the ground-truth region R_a is not in accordance with R_b, it is possible that the algorithm might detect association regions not belonging to R_a. Hence both SDR and SNR can only be used as references, and alternatively a better way to evaluate performance is to compare the input data (R_b) and the detection result (R_r) directly.

Although ideally, morphology-function analysis would maximize SDR and SNR, if the data **D** contain redundant variables (for example two variables, one in R_a and the other in $R_{\bar{a}}$, that have the same state for each case), there is no way to distinguish them without spatial information. We can expect SNR>1 if R_a is in concordance with R_b (i. e., R_a is very accurate), in which case it is possible to derive the theoretical SDR and SNR based on Bayes' theorem. Furthermore, we can claim that the morphology-function analysis algorithm performs well if both SDR and SNR are close to their corresponding theoretically maximal values. However, in this paper we do not have an accurate R_a for our data sets (although it is still very interesting to compare R_a with R_r), hence we omit the derivation of the theoretical SDR and SNR.

In addition, we can perform Receiver Operating Characteristic (ROC) curve analysis [27-28], which involves computing the True Positive Rate (TPR) and the False Positive Rate (FPR) while varying algorithm parameters, such as the BT threshold. TPR indicates the sensitivity of the method, and the False Negative Rate (FNR=1−FPR) indicates the specificity of the method. We can write TPR and FPR in terms of SDR and SNR:

$$TPR = \frac{\Omega(R_r \cap R_a)}{\Omega(R_a)} = SDR \tag{16}$$

$$FPR = \frac{\Omega(R_r \cap R_{\bar{a}})}{\Omega(R_{\bar{a}})} = \frac{\Omega(R_a)}{\Omega(R_{\bar{a}})} \cdot \frac{SDR}{SNR} \tag{17}$$

Thus, the ROC curve analysis is equivalent to the SDR-SNR analysis. Thus we only present the ROC curves in the following experimental sections.

4. Experiments for Linear Detection

This Section addresses the problem of detection of a linear morphology-function association: a subject has a functional deficit whenever there is significant atrophy in the corresponding binary-map regions.

A. Data

For this experiment we used a set of simulated cerebral-atrophy MR images based on T1-weighted gradient-echo SPGR images of 11 normal elderly subjects (average age is 70.1 years, standard deviation 5.9). We selected two gyri, the right precentral gyrus (PCG) and the left superior temporal gyrus (STG), in all subjects (both gyri were manually defined using the DISPLAY software package distributed by the Brain Imaging Center, Montreal Neurological Institute). We then introduced a 30% uniform contraction into the labeled gyri, and created 11 additional images with localized atrophy in these gyri. We term the labeled region of each subject as the atrophy-mask, or "a-mask". For each subject, we call the image without atrophy the t_1 image and the image with simulated atrophy the t_2 image. These simulated data are similar to those expected in a longitudinal study, because each pair of images (t_1 and t_2) belongs to the same subject, the only difference between the two being localized atrophy. These 22 images, as well as the corresponding a-masks, have also been used in [12].

We registered these 22 3D images with the STAR algorithm, generating RAVENS maps. The size of each RAVENS map is 256×256×129. The voxel resolution for each spatial dimension is 0.9375mm. As explained in Section 2.A, the 3D elastic warping transform in STAR algorithm preserves the brain mass of each image; therefore, the atrophic regions in the t_2 images have smaller mean intensity than those for the t_1 images. Due to the physical limitation of our computer facility, we down-sampled each image by a factor of 2, and cropped it with the largest brain-region bounding box across all images. Each of the smaller images (size=74×91×65) contains $\Omega(R_b)$=231091 brain voxels. For each subject, we repeated the warping, down-sampling and cropping procedure

(with the same parameters) on the corresponding a-mask and obtained 11 RAVENS maps of a-masks (we still call them a-masks for convenience). These a-masks are slightly different from each other; therefore, we superimposed these a-masks and binarized with the threshold 5.5 (=11/2) to generate a binary "ground truth" atrophy location R_a.

To correct registration errors, we applied an isotropic Gaussian-smoothing kernel to these images, as is customary in voxel-based morphometry [1, 14-15, 35]. We applied the same smoothing kernel to R_a, in order to generate a rational ground truth for the results obtained from a statistical analysis applied to smoothed images. Since the optimal diameter, in the sense of maximal atrophy detection, found for this data set is 9mm [12], we only show the experimental results using a 9mm-diameter smoothing kernel below. In this context, Ω(non-smoothed R_a)=3888 < Ω(smoothed R_a)=10670). Of note, R_a is the thresholded mean value of the original a-masks, however some brain regions other than R_a will also have significant intensity differences. Hence, for this data set, neither the non-smoothed R_a nor the smoothed R_a is completely accurate, although they do indicate the locations of greatest morphological change.

We then applied preprocessing Steps 2 and 3 of Fig.1 (i.e. subtraction and binarization) to the smoothed images to generate 11 binary maps for which the corresponding FV indicate a functional deficit. Notably the null hypothesis in PT (i.e. paired t-test) method indicates that when there is no atrophy, the subject has no functional deficit. Hence we created another 11 "normal" (i.e., no atrophy in t_2 images) binary maps where all cases of VVs and FV are 0. Besides this primary theoretical consideration, another practical reason for these zero maps is due to the Gaussian smoothing. Because the Gaussian-smoothing operation includes a large region around each voxel (i.e., 9mm kernel diameter relative to 0.9375mm voxel size), the value of a smoothed voxel is calculated based on the values of hundreds of neighboring voxels. If there is no atrophy in t_2 images, we would expect that after smoothing, a given voxel's intensity in the t_1 and t_2 images will be approximately equivalent, especially in the practical implementation when both input and output images of the smoothing operation are 8-bit and the smoothing itself is computed with floating-point numbers (this truncation error in data-type conversion is common in many available packages, such as SPM99). Therefore, after subtraction and binarization, most VV values are 0.

The entire data set, with 22 simulated binary maps, consisting of 11 abnormal cases for which FV=1 and 11 normal cases for which FV=0, were used as the input to the BMA algorithm.

B. Results

We compared our Bayesian methods with the standard PT method (provided in SPM99). For simplicity we call our methods BT and BC, although they have the common heuristic algorithm in Fig.3.

Fig.7 Detection results of PT, BT, and BC versus the input data for the linear morphology-function associations. (a)(b) Intensity plots of the average binary maps at the labeled two gyri, where atrophy is applied. Pixel intensity is proportional to the summation of binary maps, where 1 stands for volume loss and 0 for no volume loss. (c) The ground truth for the STG. The colored region is the smoothed R_a, which is overlaid on a subject's image (gray) for better visualization. (d)(e)(f) Detection results for the three methods evaluated.

In Fig.7(a) and (b), the intensity plots of the average binary maps at the simulated atrophy locations are shown. Comparing the smoothed R_a (for STG) in Fig.7(c) with Fig.7(b), we see that there are false-positive regions outside R_a and false-negative regions inside R_a.

In Fig.7(c)-(f), we show atrophy-detection results versus the ground truth R_a indicated by the STG atrophy-FV deficit associations. The parameters used here are p_{BT}=0.8 for BT, initial number of clusters r_L=7 and the maximum number of iterations = 100 for BC (Note that in each step of BC, although r_L was initialized to 7, only the voxels with the same label as the current RV were returned by the algorithm.), and a significance threshold t_{PT}=5 for the PT t-statistic map. The colored region in Fig.7 (c) is the R_a for STG atrophy, below which the grayscale image corresponds to a randomly chosen subject's brain, as an anatomical reference. Comparing Fig.7 (d), (e) and (f) with Fig.7(b), we see that both BT (Fig.7(e)) and BC (Fig.7(f)) correctly identify most atrophy voxels, while PT (Fig.7(d)) finds only a subset of R_a.

As expected, the results of BT and BC overlap, except for a few noisy regions in BC. This noise exists partly because BC is based on Monte-Carlo iterations, so a few pseudo-cases of the latent variable may change states even when the maximum number of iterations has been reached. In addition, we have used a large p_{BT} for BT. In

fact, BT and BC give almost identical results when p_{BT} is lowered, for example, in Fig.8 (b) the results of BT at p_{BT}=0.6 and BC at r_L=7 are very similar.

Fig.8 ROC curves of the three methods

In Fig.8 (a) and (c) we show ROC curves for both non-smoothed and smoothed R_a, respectively. Fig.8 (b) and (d) show the portions of the corresponding ROC curves in Fig.8 (a) and (c) for which we have restricted the analysis to reasonable parameters of each method, i. e., parameter values that would be used in practice. As seen in Fig.8 (b) and (d), for comparable FPRs, BT is always more sensitive than PT. Furthermore, although BC always has a higher FPR than BT and PT, BC detects more atrophy locations. When the p_{BT} is lowered to 0.6, BT produces very similar results to those of BC. However, for BT we do not know the optimal threshold beforehand, whereas BC does not depend on a user-defined threshold. Our current implementation of BC requires a value for r_L, however we could automate this choice, because when r_L is large enough (for example, >4 for these experiments), the clustered results are similar in each trial. The difference in the FPR for BC for different r_L is mainly due to a too-small r_L, which forces incorrect clustering, whereas large r_L distributes any errors across clusters. As for PT, as the threshold t_{PT} is changed to yield higher TPRs in Fig.8 (b) and (d), the corresponding parameter t_{PT} is actually unacceptable (t_{PT}=3 is too small to reflect significant variations in the t-map) for realistic applications. However, when t_{PT} is set to values typically used in practice (e. g., 10), only a small fraction of atrophy voxels (<10%) are detected. Thus, cases in which PT has low FPR correspond to low TPR as well. For example, when t_{PT}=15, the corresponding TPR is 0.0034, which means that only about 0.0034×Ω(smoothed R_a)≈36 voxels are detected. In addition, PT shares the BT's drawback of requiring a user-set threshold.

5. Experiments for Nonlinear Detection

In this section we further evaluate these methods using a more difficult problem, in which only atrophy in a

nonlinear combination of specific locations leads to a functional deficit. This structure-function relationship is one that cannot be captured by standard linear statistical tests, and it is used to demonstrate the main strength of our approach.

A. Data

For this study, we used 168 T1-weighted SPGR images of 84 normal elderly subjects. These subjects have different degrees of atrophy. For each subject, there are two images that were scanned with a 5-year interval between time t_1 and time t_2. We manually segmented these images and obtained a ventricle mask for each image. For the sake of experiment, we termed the smaller ventricle image of a subject as the t_1 ventricle and the larger ventricle image as the t_2 ventricle. That is, we first arranged the data to make all t_2 ventricles larger than the corresponding t_1 ventricles. Then we normalized these ventricle images using the STAR algorithm, and obtained two RAVENS maps for each subject. Because atrophy manifested as ventricular enlargement, the t_2 RAVENS maps have higher values than the respective t_1 maps. Then left and right ventricles were defined via a vertical line placed in the spatially normalized RAVENS maps.

Our aim in this experiment was to evaluate the performance of the BC, BT, and TT (i.e. standard t-test) methods on the following nonlinear detection problem: the functional deficit has associations with both (i.e. left and right) enlarged lateral ventricles, however only the left lateral ventricle enlargement is linearly associated with the functional deficit. The right lateral ventricle enlargement has no linear association with the functional deficit FV, i.e. p(enlarged right lateral ventricle | FV=1)=0.5 and p(non-enlarged right lateral ventricle | FV=0)=0.5. Since we already made all t_2 ventricular RAVENS maps have higher values than the corresponding t_1 maps, we further constructed 8 groups of images that displayed different patterns of atrophy, as explained next:

Table 1. Scheme to arrange the nonlinear association data set, where atrophy on both lateral ventricles is associated with the functional deficit, however only the left lateral ventricular atrophy exhibits a linear association.

Group	Number of Subjects	Pattern Name	Pattern: Enlarged Lateral Ventricle?		Functional Deficit?	Percent of Abnormal Subjects in This Pattern	Other Statistics
			Left (L)	Right (R)			
1	13	P00	0	0	0	p(FV=1 | P00) = 16/(13+16) = 55.17%	p(FV=1) = 0.5 p(L=1 | FV=1) = 0.5 p(L=1 | FV=0) = 0.19 p(R=1 | FV=1) = 0.5 p(R=1 | FV=0) = 0.5 p(FV=1 | L=1) = 0.68 p(FV=1 | L=0) = 0.40 p(FV=1 | R=1) = 0.5 p(FV=1 | R=0) = 0.5
2	16				1		
3	8	P10	1	0	0	p(FV=1 | P10) = 5/(8+5) = 38.46%	
4	5				1		
5	19	P01	0	1	0	p(FV=1 | P01) = 5/(19+5) = 20.83%	
6	5				1		
7	2	P11	1	1	0	p(FV=1 | P11) = 16/(2+16) = 88.89%	
8	16				1		

(1) We generated four ventricular atrophy patterns P00, P10, P01, and P11, which stand for no atrophy on both lateral ventricles, only left lateral ventricular atrophy, only right lateral ventricular atrophy, and both lateral ventricular atrophy, respectively. These patterns are shown in the third to fifth columns of Table 1. For simplicity, here (as in Table 1) we used shorthand L for left lateral ventricle, R for right lateral ventricle, 1 for atrophic lateral ventricle and 0 for non-atrophic lateral ventricle. Since we refer to "atrophy" as that the t_2 ventricular RAVENS map has higher value than the corresponding t_1 ventricular RAVENS map, to create the non-atrophic lateral ventricle, we swapped the t_1 and t_2 lateral ventricular RAVENS maps to make the t_1 map have higher value.

(2) We randomly divided the 84 subjects into 8 groups, each having the number of subjects and the pattern of enlarged lateral ventricles according to Table 1. For example the group 3 has 8 subjects, who have pattern P10, i.e. enlarged left lateral ventricle (i.e. L=1) and non-enlarged right lateral ventricle (i.e. R=0).

(3) For each group of subjects, we swapped the t_1 and t_2 ventricular RAVENS maps according to the patterns of enlarged ventricle listed in the fourth and fifth columns of Table 1 and set the respective functional

deficit FV states according to the sixth column of Table 1. When the lateral ventricle enlargement takes state 0 ('no'), a swap was needed. When the lateral ventricle enlargement takes state 1 ('yes'), no swap was performed. For example, for both lateral ventricles in the first group (with 13 subjects), we swapped the t_1 and t_2 ventricular RAVENS maps so that t_2 maps ended up having smaller values to those for the corresponding t_1 maps, for this group. Thus, the first group displayed no atrophy. The functional deficit FV of this group was set as 0. Another example is the 6th group (with 5 subjects), where we only swapped the left lateral t_1 and t_2 ventricular RAVENS maps and kept the right lateral ventricles unchanged. The resulting group of data displayed right-sided cerebral atrophy. We set the functional deficit FV of this group as 1.

(4) From the statistics listed in the seventh and eighth columns in Table 1, we actually designed a special simulated data set in which both lateral ventricles have strong associations with the functional deficit. Essentially, when both lateral ventricles have atrophy, i.e. pattern P11, the functional deficit is very possible, i.e. $p(\text{FV}=1 \mid \text{P11}) = 88.89\%$, although one-side ventricular atrophy has weak association with functional deficits, i.e. $p(\text{FV}=1 \mid \text{P10}) = 38.46\%$ and $p(\text{FV}=1 \mid \text{P01}) = 20.83\%$. However, when only lateral ventricle data are examined, there is no linear association between right lateral ventricular atrophy and the functional deficit, i.e. $p(\text{FV}=1 \mid \text{R}=1) = 0.5$ and $p(\text{FV}=1 \mid \text{R}=0) = 0.5$; whereas the left lateral ventricular atrophy has a linear association with the functional deficit.

Subsequently, we obtained 84 binary maps after the subtraction (we used t_2–t_1 here, in contrast to the experiment in last section) and binarization steps in Fig.1. According to the FV setting in Table 1, we have 42 normal subject (without atrophy) and 42 abnormal subjects (with atrophy). The functional deficit of these atrophic subjects arises under specific combinations of bilateral atrophy. The average binary map for these 84 subjects is shown in Fig.9 (a).

B. Results

We applied BT, BC, and TT to these data. Via the design of this experiment, we knew that both lateral ventricles are associated with the functional deficit. However, the right lateral ventricle is not linearly associated with FV. Hence a standard linear statistical test tool, e.g. TT, can hardly detect the right lateral ventricular atrophy, although it was expected to find the left lateral ventricular atrophy. Fig.9 (b) shows the result of the TT method, where only the left lateral ventricle shows up on the thresholded p-map (with the threshold p-value = 0.05). In contrast, in our BMA results in Fig.9 (c) and (d) (with parameters p_{BT}=0.8 and r_L=3), both BT and BC detect the morphology-function associations of both lateral ventricles. These results confirm our expectation that BT and BC can detect nonlinear morphology-function associations.

In BC result shown in Fig.9 (d), the left lateral ventricle does not appear as bright as that of the BT result in Fig.9 (c). This is because of the imperfect convergence of the Monte Carlo iteration in BC. Despite this point, the results of BT and BC accord well with each other.

Fig.9 Detection results of TT, BT, and BC versus the average binary map for the nonlinear morphology-function associations. The red and white regions correspond to two clusters of associations found by BT and BC methods. The result of TT is painted as red because it corresponds the red region in results of BT and BC.

6. Discussion

The proposed BMA methods are generally applicable to many similar problems, besides detection of morphology-function associations. This advantage makes BMA a potentially powerful tool in discovering interesting morphological characteristics that are associated with non-image categorical or discretized variables. Although we only present the experiments on binary variables, BMA can also analyze multi-valued discrete variables. Further, in case there are two or more FVs, our BMA method would return a large Bayesian network, in which a subnetwork is generated for each FV.

In most applications the number of RV classes is less than 10, for example in Section 4 we obtained 1 RV and

in Section 5 we obtained 2 major RVs; the largest number of RV classes we have ever obtained is 8 in other experiments [29]. With this result in mind, regarding the algorithm in Fig.3, we observed that when there are fewer than 10 RVs, the computational complexity of the heuristic search of network structures is approximately linear in the total number of voxels in the reported experiments. Therefore, BMA methods may be scalable to millions of voxels, although BC is slow due to its reliance on the Monte Carlo algorithm. However, in our experiments regarding detection of atrophy-function associations, we found that BC often converges rapidly, usually within 20 iterations.

In our experimental design, only binary variables were considered. This setting offers the advantage that voxel variables have clear meaning of their values, i.e. 1 for volume contraction and 0 for non-contraction. However, it is possible to use 3 states for each voxel variable, i.e. 2 for volume contraction, 1 for volume expansion, and 0 for no change. The latter setting is useful to detect more general brain morphological changes, besides the atrophy. Nonetheless for the purpose of atrophy detection, the binary variables are sufficient, and avoid confusion in explaining experimental results.

For the data in Section 4, in which we used a 9mm diameter Gaussian smoothing kernel, the intensity difference between smoothed t_1 and t_2 voxels was very small, typically within the range [-10,10]. Hence, the naïve binarization threshold 0 appeared to be necessary, because a threshold larger than 0 would have caused too much signal loss and thus very low TPR. One solution would be to increase the contrast of the comparison maps. For example, we could voxel-wise divide the t_2 image by the t_1 image; this operation would produce ratio maps with much higher contrast than difference maps. However in the case we were successful with the difference maps, it appears unnecessary to enlarge the computation load by using ratio maps, although higher contrast (e. g., from clustering methods based on the Expectation Maximization algorithm) would allow us to manually or automatically choose a better binarization threshold to maximize theoretical SNR and/or SDR. Furthermore, in those cases for which we have better ground truth, we can calculate the best SNR and/or SDR for many different Gaussian kernel diameters to find the optimal kernel.

We plan the following major areas of further development:

- Spatial information: when spatial information is integrated into the algorithm, some noisy VVs can be

removed. For example, an isolated bright point on the average binary map is usually meaningless. Therefore, it is reasonable to eliminate such voxels using neighborhood filtering, or probability distributions over the spatial distribution of regions associated with functional variables.

- Number of cases: Although we omit the derivation here, analytically the number of cases will play an important role in improving both SDR and SNR (i. e., increasing TPR and decreasing FPR).
- Different metric and model selection heuristics: Although we employ the Bayesian metric in eqn.(5) and propose the heuristic method in Fig.3, the proposed BMA paradigm itself is independent of particular Bayesian network metrics and heuristic model selection methods. We plan to investigate several others approaches described in the Bayesian network learning literature [19, 22-23, 38, 6].

The main limitation of the BMA approach is that variables are required to be discrete. Fortunately, in many applications the data can be discretized. Also, it is possible to extend our work to continuous (e.g. Gaussian) variables (for a review of continuous-variable Bayesian network, see [36]).

7. Conclusion

In this paper we have described a framework for morphology-function analysis, based on a Bayesian-network model of relationships among image and functional variables. The algorithms based on this framework generate sets of voxels whose members have similar probabilistic associations with the functional variable(s). Two methods implemented within this framework, Bayesian thresholding and Bayesian clustering, are examples of how this framework can be used to generate equivalent-voxel sets. The Bayesian thresholding method is simpler and faster, however it requires a predefined threshold for determining whether two voxels have similar conditional probability distributions given the functional variable. The Bayesian clustering method utilizes a latent-variable Bayesian-network model, and the Monte-Carlo algorithm, to generate equivalent-voxel sets. Bayesian clustering takes longer time than Bayesian thresholding, however the former does not require a user-defined threshold, which is the principal limitation of the latter.

We compared these Bayesian methods to the standard statistical tests (i.e. t-test and paired t-test) for both linear and nonlinear morphology-function-detection problems. Our methods succeeded in both cases, whereas the paired t-test detected the atrophy region for the linear-association problem, and the t-test failed to detect nonlinear associations. It is possible to extend our framework to effectively detect other morphology-function associations between image and categorical variables.

Acknowledgement

This work was supported by The Human Brain Project, National Institutes of Health grant R01 AG13743, which is funded by the National Institute of Aging, the National Institute of Mental Health, the National Aeronautics and Space Administration, and the National Cancer Institute. Dr. Herskovits was also supported by a Richard S. Ross Clinician Scientist Award. We thank David Chickering for assistance with latent-variable Bayesian clustering methods.

References

[1] Ashburner, J. and Friston, K.J., "Voxel-based morphometry: the methods," NeuroImage, vol.11, pp.805-821, 2000.

[2] Bookstein, F.L., "Principal warps: thin-plate splines and the decomposition of deformations," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.11, pp.567-585, 1989.

[3] Bouckaert, R.R., "Probabilistic network construction using the minimum description length principle," Technical Report, RUU-CS-94-27, Dept of Computer Science, Utrecht University, 1994.

[4] Cheeseman, P., and Stutz, J., "Bayesian classification (Autoclass): theory and results," In Fayyad, U., Piatesky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Eds.), Advances In Knowledge Discovery And Data Mining, pp.153-180. Menlo Park, CA: AAAI Press, 1995.

[5] Chib, S., "Marginal likelihood from the Gibbs output," Journal of the American Statistical Association, vol.90, pp.1313-1321, 1995.

[6] Chickering, D.M., and Heckerman, D., "Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables," Machine Learning, vol.29, pp.181-212, 1997.

[7] Clogg, C., "Latent class models," In Arminger, G., Clogg, C., & Sobel, M. (Eds.), Handbook Of Statistical Modeling For The Social And Behavioral Sciences. Plenum Press, New York., 1995.

[8] Cooper, G., and Herskovits, E., "A Bayesian method for the induction of probabilistic networks from data," Machine Learning, vol.9, pp.309-347, 1992.

[9] Davatzikos, C, "Spatial transformation and registration of brain images using deformable models," Computer Vision and Image Understanding, vol.66, no.2, pp.207-222, 1997.

[10] Davatzikos, C., "Mapping of image data to stereotaxic spaces: applications to brain mapping," Human Brain Mapping, vol.6, pp.334-338, 1998.

[11] Davatzikos, C., "Spatial normalization of 3D brain images using deformable models," Journal of Computer Assisted Tomography, vol.20, no.4, pp.656-665, 1996.

[12] Davatzikos, C., Genc, A., Xu, D.R., and Resnick, R.M., "Voxel-based morphometry using RAVENS maps: methods and validation using simulated longitudinal atrophy," NeuroImage, vol.14, pp.1361-1369, 2001.

[13] Davatzikos, C., Vaillant, M., Resnick, S., Prince, J.L., Letovsky, S., and Bryan, R.N., "A computerized method for morphological analysis of the corpus callosum", Journal of Computer Assisted Tomography, vol. 20, pp. 88-97, Jan./Feb. 1996.

[14] Friston, K.J., Ashburner, J., Frith, C.D., Poline, J.B., Heather, J.D., and Frackowiak, R.S.J., "Spatial registration and normalization of images," Human Brain Mapping, vol.2, pp.165-189, 1995.

[15] Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith C.D., and Frackowiak, R.S.J., "Statistical parametric maps in functional imaging: a general linear approach", Human Brain Mapping, vol.2, pp.189-210, 1995.

[16] Gaser, C., Volz, H.P., Kiebel, S., Riehemann, S., and Sauer, H., "Detecting structural changes in whole brain based on nonlinear deformations-application to schizophrenia research," NeuroImage, vol.10, pp.107-113, 1999.

[17] Gee, J.C., Reivich, M., and Bajcsy, R., "Elastically Deforming 3D Atlas to Match Anatomical Brain Images", Journal of Computer Assisted Tomography, vol.17, pp.225-236, 1993.

[18] Geman, S., and Geman, D. "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," IEEE Transactions on Pattern Analysis And Machine Intelligence, vol.6, pp.721-742, 1984.

[19] Glymour, C., and Cooper, G.F., (ed.), Computation, Causation, and Discovery, AAAI/MIT Press, 1999.

[20] Goetz, C.G., and Pappert E.J., Textbook Of Clinical Neurology, (1st Ed.), W.B. Saunders Company, 1999.

[21] Goldszal, A.F., Davatzikos, C., Pham, D.L., Yan, X.H., Bryan, R.N., and Resnick, S.M, "An image processing system for qualitative and quantitative volumetric analysis of brain images," Journal of Computer Assisted Tomography, vol.22, no.5, pp.827-837, 1998.

[22] Heckerman D, "Bayesian Networks For Data Mining," Data Mining and Knowledge Discovery, vol.1, no.1, pp.79-119, 1997.

[23] Heckerman, D., Geiger, D., and Chickering, D., "Learning Bayesian networks: the combination of knowledge and statistical data," Machine Learning, vol.20, pp.197-243, 1995.

[24] Herskovits, E.H., "Computer-based probabilistic-network construction," Doctoral Dissertation, Medical Informatics, Stanford University, 1991.

[25] Jensen, F.V., An Introduction To Bayesian Networks, Springer Press, 1996

[26] McIntosh, AR, Bookstein, FL, Haxby, JV & Grady, CL, (1996). Spatial pattern analysis of functional brain images using partial least squares. NeuroImage, vol.3, 143-157.

[27] Metz, C.E. "Basic principles of ROC analysis," Seminars In Nuclear Medicine, vol.8, pp.283-98, 1978

[28] Metz, C.E., "ROC methodology in radiologic imaging," Investigative Radiology, vol.21, pp.720-733, 1986.

[29] Peng, H.C., and Long, F.H., "A Bayesian learning algorithm of discrete variables for automatically mining irregular features of pattern images," the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001) in conjunction with ACM SIG/KDD2001, San Francisco, USA, Aug, 2001.

[30] Raftery, A., Hypothesis Testing and Model Selection, Chap.10, Chapman and Hall, 1996.

[31] Spiegelhalter, D., Dawid, A., Lauritzen, S., and Cowell, R., "Bayesian analysis in expert systems," Statistical Science, vol.8, pp.219-282, 1993.

[32] Thompson, P.M., and Toga, A.W., "A surface-based technique for warping three-dimensional images of the brain," IEEE Transactions on Medical Imaging, vol.15, pp.402-417, 1996.

[33] Thompson, P.M., and Toga, A.W., "Detection, visualization and animation of abnormal anatomic structure with a deformable probabilistic brain atlas based on random vector field transformations," Medical Image Analysis, vol.1, no.4, pp.271-294, 1997.

[34] Thompson, P.M., MacDonald, D., Mega, M.S., Holmes, C.J., Evans, A.C., and Toga, A.W., "Detection and mapping of abnormal brain structure with a probabilistic atlas of cortical surfaces," Journal of Computer Assisted Tomography,

vol.21, pp.567-581, 1997.

[35] Woermann, F.G., Free, S.L., Koepp, M.J., Ashburner, J., and Duncan, J.S., "Voxel-by-voxel comparison of automatically segmented cerebral gray matter – a rater-independent comparison of structural MRI in patients with epilepsy," NeuroImage, vol.10, pp.373-384, 1999.

[36] Roweis, S., and Ghahramani, Z., "A unifying review of linear Gaussian models," Neural Computation, vol.11, no.2, pp.305-345, 1999.

[37] Shen, D.G., Herskovits, E.H., and Davatzikos, "An adaptive-focus statistical shape model for segmentation and shape modeling of 3D brain images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.20, no.4, pp.257-270, 2001.

[38] Buntine, W, "A guide to the literature on learning probabilistic networks from data," IEEE Transactions on Knowledge And Data Engineering, vol.8, no.2, pp.195-210, 1996.

[39] Blalock, H., Causal Inferences in Nonexperimental Research, Chapel Hill, NC: UNC Press, 1961.

A Bayesian Morphometry Algorithm

Hanchuan Peng *, Edward Herskovits, and Christos Davatzikos

Center for Biomedical Image Computing,

Department of Radiology, School of Medicine, Johns Hopkins University,

601 N Caroline St, JHOC 3230, Baltimore, MD, 21287, USA.

Email: phc@cbmv.jhu.edu, ehh@braid.rad.jhu.edu, hristos@rad.jhu.edu

List of Figures

Figures

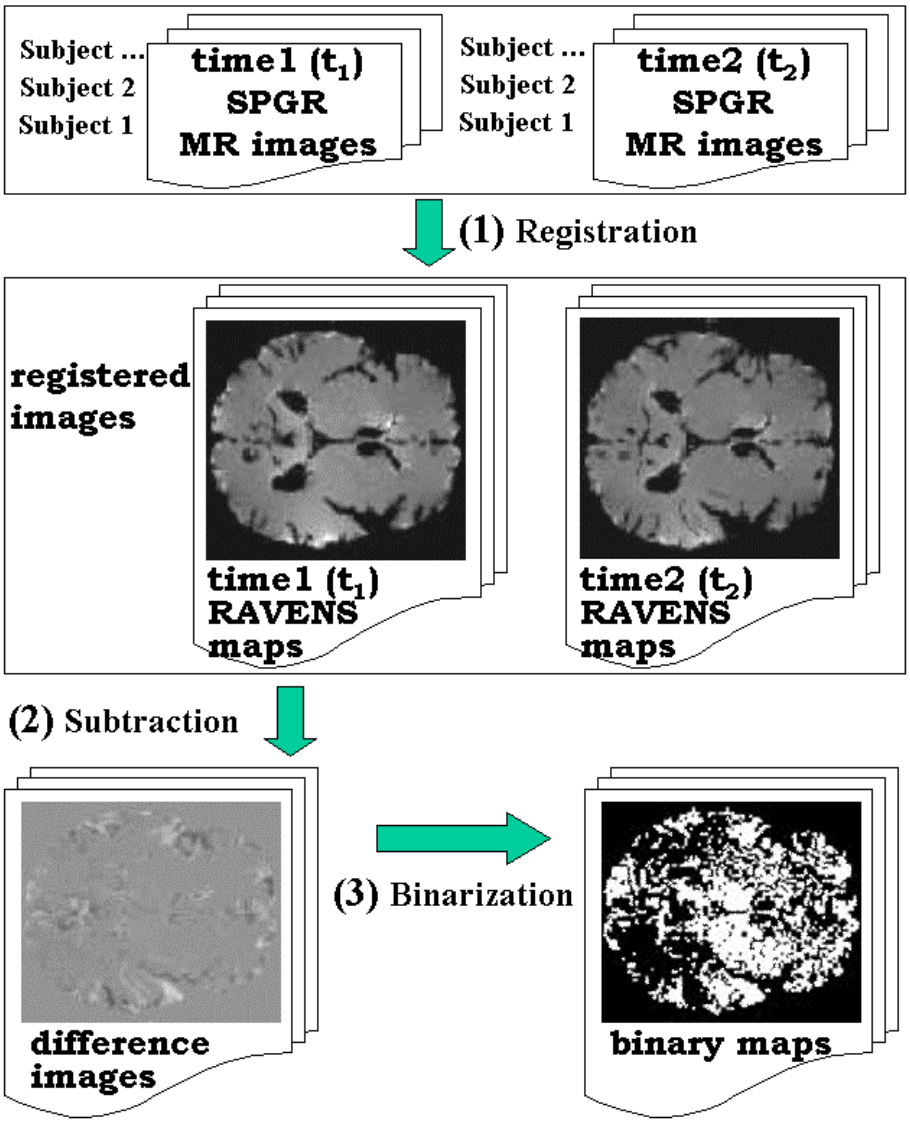


Fig.1 Three major image preprocessing steps, i.e. registration, subtraction and thresholding (binarization)

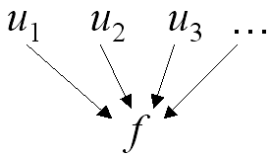


Fig.2 The general Bayesian-network structure for representing morphology-function associations

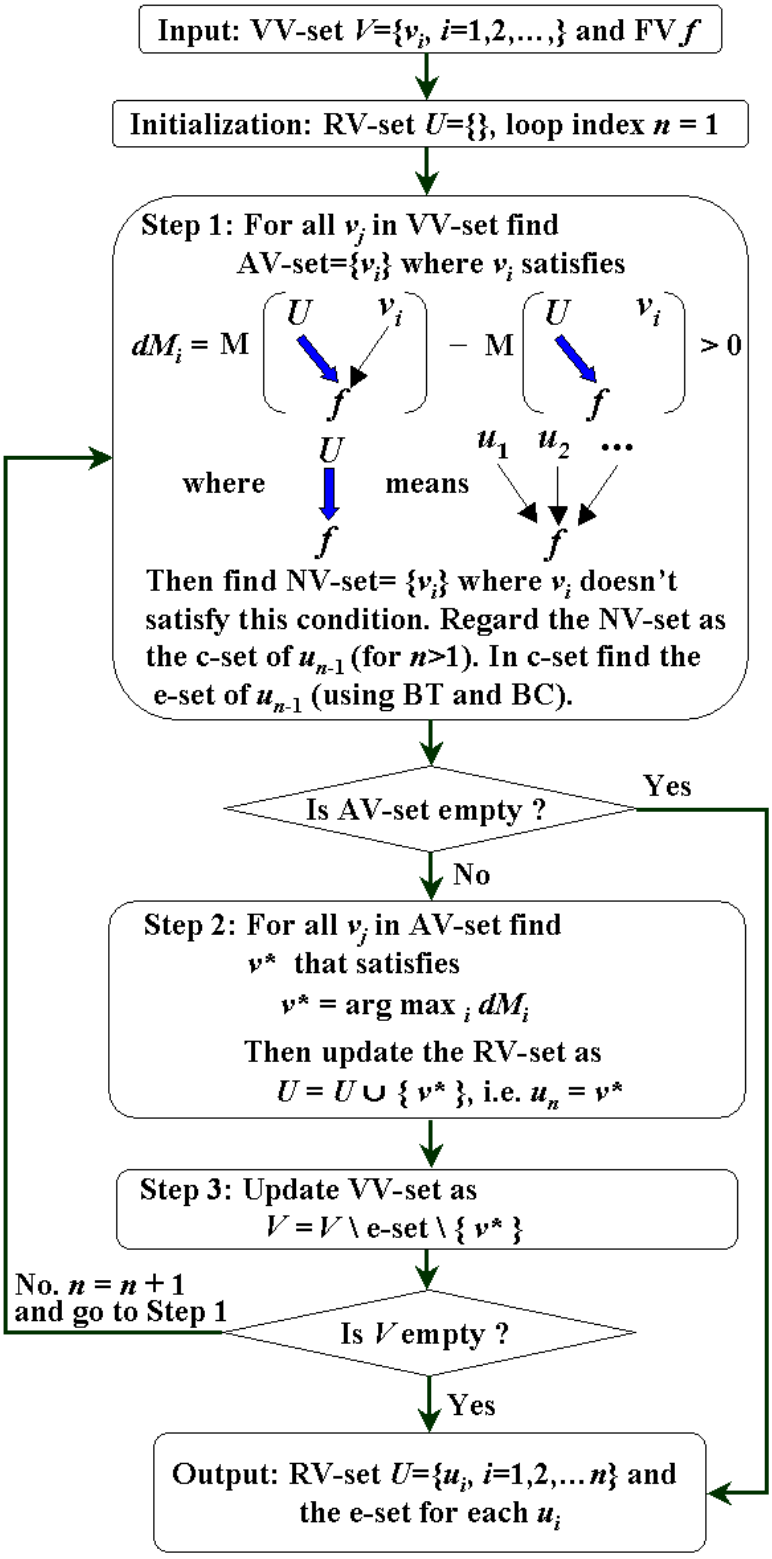


Fig.3 Flowchart of the algorithm for RV-set generation

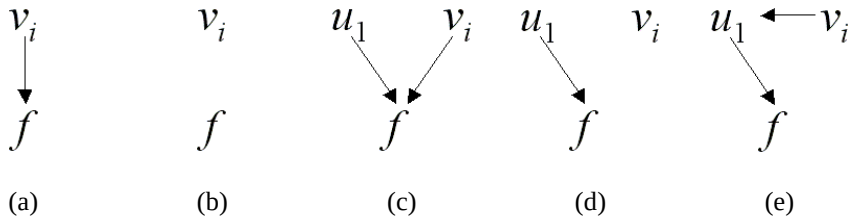


Fig.4 Alternative structure-function Bayesian-network models

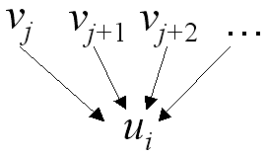


Fig.5 The general Bayesian-network structure for the BT method

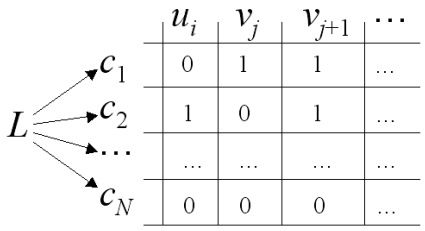
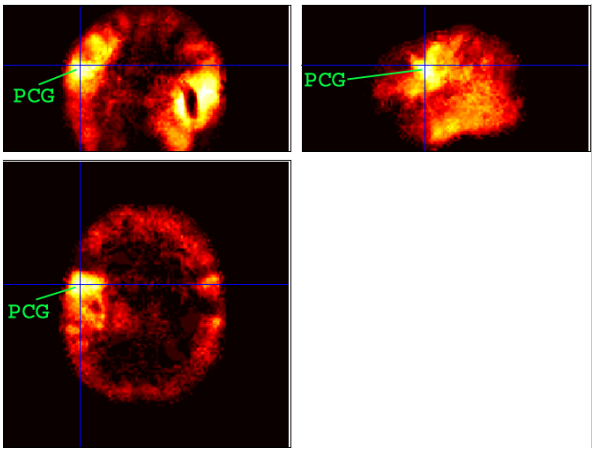
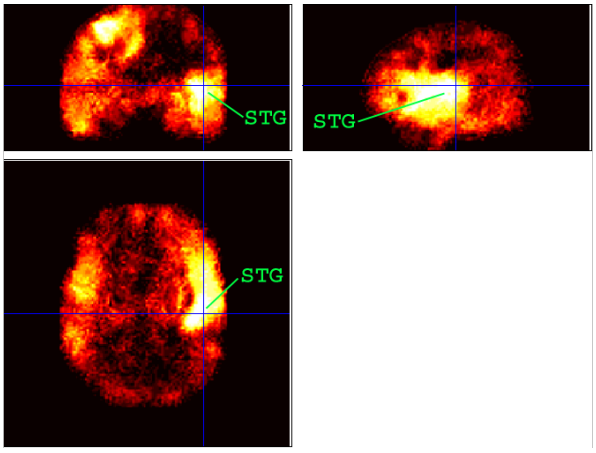


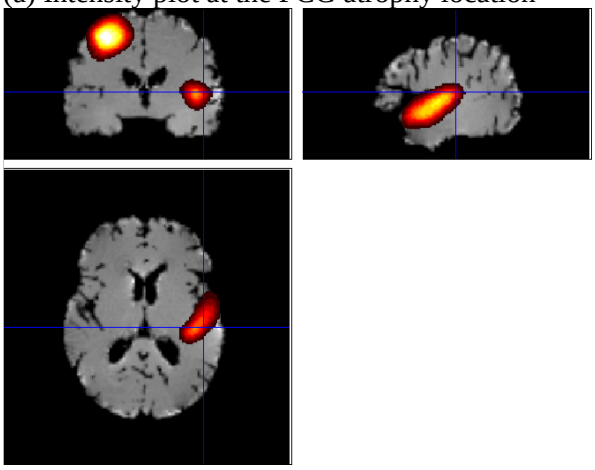
Fig.6 The latent-variable Bayesian-network structure for the BC method



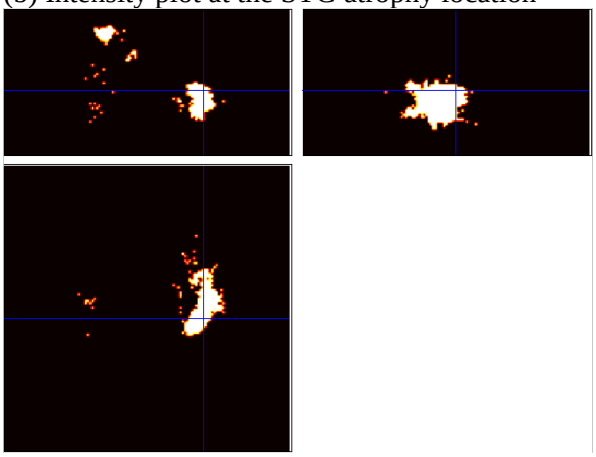
(a) Intensity plot at the PCG atrophy location



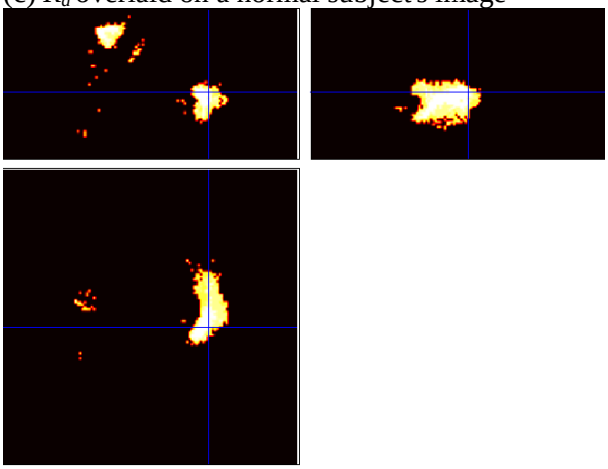
(b) Intensity plot at the STG atrophy location



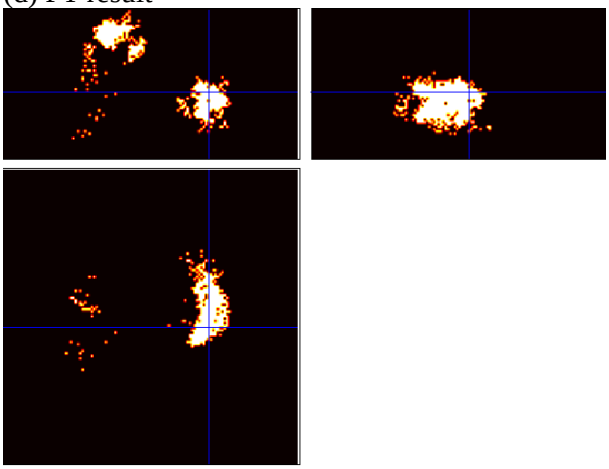
(c) R_a overlaid on a normal subject's image



(d) PT result

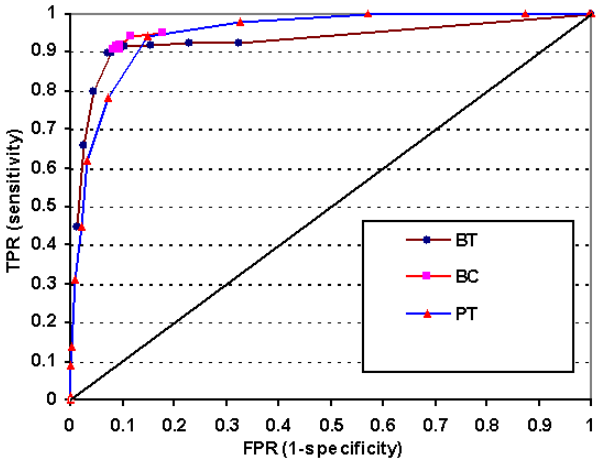


(e) BT result

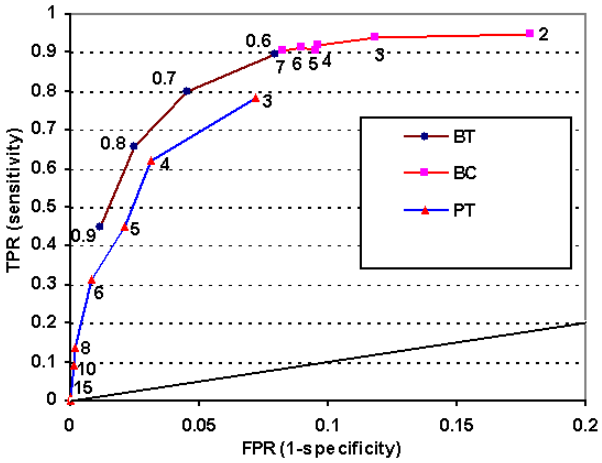


(f) BC result

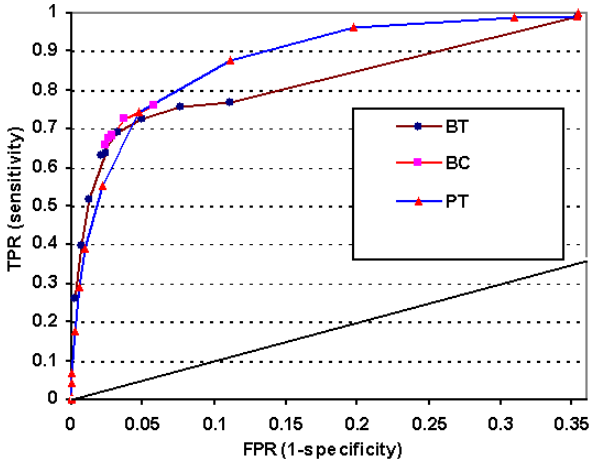
Fig.7 Detection results of PT, BT, and BC versus the input data for the linear morphology-function associations. (a)(b) Intensity plots of the average binary maps at the labeled two gyri, where atrophy is applied. Pixel intensity is proportional to the summation of binary maps, where 1 stands for volume loss and 0 for no volume loss. (c) The ground truth for the STG. The colored region is the smoothed R_a, which is overlaid on a subject's image (gray) for better visualization. (d)(e)(f) Detection results for the three methods evaluated.



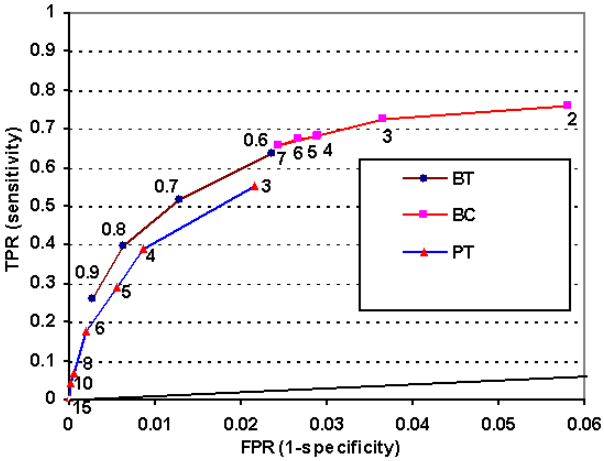
(a) ROC curve (non-smoothed R_a) for all parameters



(b) ROC curve (non-smoothed R_a) for meaningful parameters



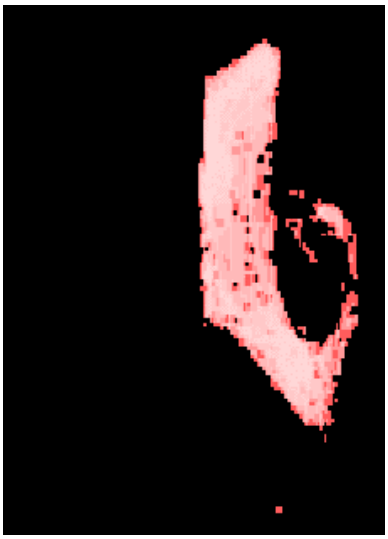
(c) ROC curve (smoothed R_a) for all parameters



(d) ROC curve (smoothed R_a) for meaningful parameters

Fig.8 ROC curves of the three methods

(a) Average binary ventricle of the 84 subjects (axial view from below)



(b) TT (p_{TT}=0.05) (axial view from below)

(c) BT (p_{BT}=0.8) (axial view from below)

(d) BC (r_L=3, loop=50) (axial view from below)

Fig.9 Detection results of TT, BT, and BC versus the average binary map for the nonlinear morphology-function associations. The red and white regions correspond to two clusters of associations found by BT and BC methods. The result of TT is painted as red because it corresponds the red region in results of BT and BC.